Virginia

Standards of Learning Assessments

**Standards of Learning (SOL) Tests
Validity
and
Reliability Information
Spring 1998 Administration**

EXECUTIVE SUMMARY

From their inception, the Virginia *Standards of Learning* (SOL) tests have been developed with the utmost attention to the technical requirements of a high-stakes testing program. Two key areas of technical merit are essential in such tests: validity and reliability. Considerations regarding test validity and reliability are present throughout the SOL test development process.

The most important criterion for judging the validity of any test is procedural and concerns the question: Are processes in place that ensure that test questions measure the content upon which the test is based? This is unequivocally the case with the Virginia SOL tests. Throughout the Content Review Committee process, coupled with review of statistical information from field test administrations, Virginia educators work with the Department of Education and the testing contractor, Harcourt Brace Educational Measurement, to ensure that every item that appears on a Virginia operational SOL test matches the SOLs and the test specifications.

A second type of validity evidence involves correlations between the SOL tests and other related measures. Review of information from other measures of student performance, the *Stanford 9* and the *Literacy Passport Test* (LPT), indicates that schools that performed well on these measures generally performed well on related SOL tests.

Test reliability information from the Spring 1998 SOL test administration, released in September 1998, confirms that Virginia's SOL tests meet accepted technical requirements for reliability of scores on high-stakes tests.

While the testing contractor and the Department of Education bring technical expertise to the development of the SOL tests, the validity and reliability statistics from the first administration of the SOL tests have been reviewed by outside testing experts and found to be solid for such tests. Testing experts with experience in state testing programs throughout the country attest to the solid foundation upon which the Virginia SOL tests rest.

Evidence of validity and reliability of test scores from the first SOL test administration should dispel any notion that students did not perform well because the tests themselves were faulty. Instead, SOL test performance supports the notion that many schools in Virginia are doing an exceptional job instructing students in the content of the SOLs, while many other schools have a considerable task ahead of them.

Ensuring Content Validity

The most important validity evidence for a statewide, criterion-referenced test is evidence that the test items measure the state objectives and that each test form matches the Virginia test specifications. Such evidence usually consists of professional judgments made by representative panels of content experts.

Dr. S. E. Phillips, Professor, Michigan State University
Regarding the Virginia SOL testing program

As noted by Dr. Phillips, Virginia's test development process employs such a procedure for the express purpose of assuring that each and every SOL test administered measures the SOLs in the content area and complies with the test specifications.

Each SOL test is developed with the cooperation of a Content Review Committee composed of Virginia educators who have experience and expertise in the content area and grade level of the test. Working with the Department and the testing contractor, the Content Review Committees review each test question before it is field tested. In order to make it to field test status in Virginia, the potential question must, in the eyes of the Content Review Committee, meet the following four criteria:

- ✓ Does the question measure the SOL it was designed to measure?
- ✓ Does the question appropriately measure content or skills that students in Virginia should be expected to learn by the spring of the designated grade level or near the end of the course?
- ✓ Is the difficulty of the question appropriate?
- ✓ Is the question free from content that stereotypes, offends, or unfairly penalizes students on the basis of personal characteristics such as gender, ethnicity, religion, or socioeconomic status?

If the question meets these criteria, it is used as a field test question with Virginia students during a regular administration of the SOL tests. Test questions used on the Spring 1998 SOL tests were field tested with Virginia students in Spring and Fall 1997. In subsequent administrations, field test questions appear among the operational questions on each SOL test. In this way a steady stream of new test questions is available for use.

Review of Statistical Information Regarding Each
Potential Test Question

After test questions have been field tested, the Content Review Committee is reconvened to review the questions and make the same four judgments again, but this time with statistics about the field-tested questions.

The following field test statistics are used to decide if questions will appear in operational test forms:

- ✓ Traditional item statistics
- ✓ Rasch item statistics
- ✓ Differential Item Functioning (DIF)

Traditional Item Statistics: frequency distributions; item mean (P values); standard deviations; n count; point-biserial correlations; biserial correlation; item reliability index; response distribution for each option for all respondents; for high, middle, and low ability groups; by gender; and by ethnic group. In addition, traditional item difficulty and item discrimination statistics are computed.

Rasch Item Statistics: To supplement the traditional statistics, item difficulty parameter estimates based on Item Response Theory (IRT) are computed. Under the Rasch model of IRT, a common underlying construct is assumed to be measurable and estimable as a function of item or test performance, making it possible to estimate item difficulty and item fit.

Differential Item Functioning:

The Rasch IRT method of computing DIF is also employed to provide item difficulty estimates among demographic groups. Under the assumptions of the Rasch model, the only reason for differences in item difficulty statistics among groups is some group characteristic other than achievement. When the Rasch item difficulty estimates are statistically significant among groups, it is an indicator that further examination is warranted. The Rasch IRT procedure was used to compare white and African-American students, white and Hispanic students, and male and female students.

DIF statistical procedures such as the Mantel-Haenszel Alpha compute the probability that one demographic group is more likely to answer an item correctly than another group. This information is useful in reviewing items and tests for potential bias. High values of the Mantel-Haenszel Alpha indicate that an item interacts differently among equally able students in the reference and comparison groups. When the probability is significantly different across groups, an item warrants further examination. The Mantel-

Haenszel Alpha procedure is used to compare white and African-American students, white and Hispanic students, and male and female students.

After review of the above statistical information **for each field-tested question**, the Content Review Committees again must answer the essential four questions:

- ✓ Does the question measure the SOL it was designed to measure?
- ✓ Does the question appropriately measure content or skills that students in Virginia should be expected to learn by the spring of the designated grade level or near the end of the course?
- ✓ Is the difficulty of the question appropriate?
- ✓ Is the question free from content that stereotypes, offends, or unfairly penalizes students on the basis of personal characteristics such as gender, ethnicity, religion, or socioeconomic status?

If the answers to these four questions are "yes," then the test question is placed in the item bank for potential use in an operational SOL test form.

Additional Attention to Bias in Test Questions

Because passing the high school SOL assessments will be a graduation requirement, it is especially important that the tests be free of factors that have an unfair impact upon a group of students. Therefore, an additional bias review is conducted by a separate Bias Review Committee representing each content area to be tested. Bias Review Committee members are asked to scrutinize items for potential stereotyping or other forms of bias. The purpose of the bias review is to identify any items that appear to have the potential to treat any ethnic, gender, or regional group of students differently from other groups. Committee members examine the response distribution for each of the demographic groups identified for the review. The intent of the review is to determine if members of a certain group were drawn to one or more of the answer choices for the item. If a large percent of one group selected a particular response or did not select a particular response, the item is carefully examined.

Training and procedures are similar to those carried out during the Item Review meetings. The Bias Review Committee's task focuses solely on reviewing test items for potential bias after the items have been reviewed by the Content Review Committees. It is the Bias Review Committee's responsibility to ensure that items are fair to all students and that all students would have an equal opportunity to demonstrate achievement regardless of gender, ethnic background, religion, socioeconomic status, or geographic region.

Constructing Forms According to Test Blueprint Specifications

Each SOL test is constructed according to the specifications of the SOL test blueprint. In preparing a test form for operational use, the testing contractor selects questions from the item bank in accordance with the blueprint specifications. Once draft test forms are constructed, the Content Review Committees are again convened. Committee members assume the task of approving or editing two forms of each grade level or end-of-course test to determine the content validity and equivalency of the test forms as a whole. While the previous committee reviews were concerned with individual questions, the focus of this review is the full operational test forms. At this stage there may be additional minor edits or revisions.

It is after this final stage that the test forms are ready for use in an SOL test administration.

Same Technical Rigor Applied to Every Item on a Virginia SOL Test

> *The development of Virginia's SOL tests mirrors the processes applied to the development of other high-stakes tests. The test development process used in Virginia is exemplary in its use of technical criteria for judging the quality of test questions, as well as the involvement of Virginia educators. The hard work of these panels has resulted in tests that are valid measures of the SOLs.*

> Dr. Joanne M. Lenke, President
> Harcourt Brace Educational Measurement

All items that appeared on the Spring 1998 SOL tests were subjected to this procedure. All items that appear on subsequent forms of the SOL tests will be developed in this same manner and receive this same level of professional and technical scrutiny. Each SOL test is grounded in a complex process that meshes technical requirements for valid tests with practical considerations from everyday teaching.

Outside Review of Procedures Conducted

The item and test development procedures follow best measurement practice for establishing content validity. Judgmental, statistical, and item option reviews provide comprehensive scrutiny of all items for potentially unfair performance effects for African-Americans and Hispanics. Only items that have passed rigorous professional and statistical scrutiny are retained for use on actual test forms. Continuous pretesting of new items under live administration conditions provides ideal data for constructing new test forms. Construction of test forms to match the Virginia SOL blueprints ensures that the specified skills are tested according to the intended plan of skill emphases.

Dr. S. E. Phillips, Professor, Michigan State University
Regarding the Virginia SOL testing program

Evidence for validity based on the content of the tests has been carefully gathered and clearly supports the inference that the test scores indicate student knowledge and skill as defined by the SOLs. An appropriate review process by content experts of individual items, as well as the tests as a whole, has been accomplished for each of the tests. The questions used in the content review process were appropriate for focusing the experts' judgments on the match between the content measured in the item and the corresponding SOL. The procedures used to ensure that the items were not biased or unfair were extensive and appropriate. The use of test blueprints in the preparation of the tests is consistent with professional testing guidelines. It is particularly noteworthy that the judges reviewd the difficulty of the items, since appropriate difficulty is key to establishing reliability and standards for reaching acceptable levels of performance.

Dr. James McMillan, Professor
Virginia Commonwealth University

The test development process that the Virginia Department of Education followed to ensure the technical adequacy of the Virginia SOL assessment program mimics that of other high-stakes testing programs across the country and reflects the Code of Fair Testing Practices in Education *(Code) as endorsed by AERA, APA, and NCME. Specifically, the selection of content knowledge and skills assessed through the Content Review Committee process ([performed by] Virginia educators) is to be noted. The professional judgments provided by this committee contribute to the evidence that the test items and forms match the state standards and test specifications.*

Dr. Tonya Moon
University of Virginia

Several individuals with expertise in the development of high-stakes tests were asked to review the validity information compiled from the Spring 1998 administration of the SOL tests. Their reviews consistently support the appropriateness of the procedures and statistical information used in the development of the SOL tests.

SOL Results Similar to Those of Other Tests

> *Another type of validity evidence that may provide useful descriptive information about a test is correlations with other measures. The other measures can be instruments that measure similar content or different content than the test of interest. For measures whose content domains overlap with the content domain of the test of interest, one might expect relatively high correlations. However, when the two instruments measure different content domains, their correlations would be expected to be relatively low. For example, the correlation between scores for two reading tests would be expected to be much higher than the correlation between scores from a reading test and a science test.*

> Dr. S. E. Phillips, Professor, Michigan State University
> Regarding the Virginia SOL testing program

One important piece of evidence of the validity of a SOL test is if the SOL test results are similar to those of other tests. For instance, it is reasonable to expect that the relative standing of schools on results of the SOL tests would be similar to the relative standing of schools on the *Stanford 9* and the *Literacy Passport Tests*. Do schools that score well on the *Stanford 9* or the *LPT* also score well on the SOL tests in content areas where there are similar knowledge and skills?

It must be acknowledged and emphasized that differences do exist between the Virginia SOL tests and the *Stanford 9* and the *LPT*. The *Stanford 9* is based on a broad representation of national curricula and is not specific to Virginia. However, in the areas of reading, language, and mathematics, there are areas of overlapping content. The *LPT* is a basic skills test anchored in previous versions of Virginia's SOLs and thus represents a less rigorous set of expectations than those of the SOL testing program.

In the content areas and grade levels where there were reasonable matches of content, school pass rates on the SOL tests have been statistically correlated with national percentile ranks on the *Stanford 9* and/or pass rates on the *LPT*. These data show a strong relationship between the relative standing of Virginia's schools on the SOL tests and both the *Stanford 9* and the *LPT*. While overall performance on the SOL tests is dramatically lower than on the *Stanford 9* and the *LPT*, the relative standing among schools is very similar. Though varying among grades and content areas, schools that scored well on the *Stanford 9* or *LPT* generally scored well on related SOL tests, and vice versa.

The Department of Education sought outside expert review of the correlation information relating the SOL test results to those of the *Stanford 9* and the *Literacy Passport Test*. Outside review confirms the support the correlation information provides to the technical quality of the SOL tests.

> *The school level rank order correlations for the Virginia and* Stanford 9 *subtests, summarized in Table 1,* ***are in the expected range****.* [emphasis added] *Based on these correlations, the* Virginia SOL *tests appear to rank order schools most similarly to the* Stanford 9 *in grade 8 and least similarly at the high school level. Correlations between the* Virginia SOL *and* Virginia LPT *are higher in language arts than mathematics. The* SOL *mathematics tests appear to rank order schools more similarly to* Stanford 9 *mathematics problem solving than mathematics procedures.*

> *Overall, approximately 28%-72% of the variance in school rankings is shared between the* Virginia SOL *and* Stanford 9 *and approximately 29%-56% between the* Virginia SOL *and* LPT *tests, with most falling nearer 50%. These values support the proposition that the three tests measure the same broad skills such as English or mathematics but differ in some of the specific content and skills measured.*

> *The substantial correlations with other measures provide supporting validity evidence for the Virginia SOL tests.*

> <div align="right">Dr. S. E. Phillips, Professor, Michigan State University
Regarding the Virginia SOL testing program</div>

> *Evidence for validity based on relations to other measures has been provided and is more than adequate for this type of test. The moderate magnitude of the correlations between the* SOL *tests and established measures of similar knowledge and skills is what would be expected since the measures that are correlated do not match each other completely. This is, not all of the content of the* SOL *tests is covered on the* Stanford 9 *or* Literacy Passport Test, *and these tests include content not included in the* SOLs. *Consequently, the correlations obtained in the range between .53 and .85 are consistent with expectations. Furthermore, the fact that correlations between measures of content with greater overlap are higher than correlations between measure of content with only moderate overlap supports the contention that the* SOL *tests represent the* SOLs *as indicated in the tables of specifications.*

> <div align="right">Dr. James McMillan, Professor
Virginia Commonwealth University</div>

Table 1
Virginia School Level Rank Order Correlations*

Comparison	Grade 3	Grade 5	Grade 8	H.S./ Grade 11
SOL English: Reading/Literature and Research & *Stanford 9* Reading	.76-.78	.76-.78	.80-.81	.57-.62
SOL English: Writing & *Stanford 9* Language	.72	.76	.82	.71
SOL Mathematics & *Stanford 9* Math		.67-.76	.67-.76	.77-.85
SOL Algebra I, II, and Geometry & *Stanford 9* Total Mathematics				.53-.71
SOL English: Reading/Literature & Research & Grade 6 *LPT* Reading		.64	.75	
SOL English: Writing & Grade 6 *LPT* Writing		.68	.61	
SOL Mathematics & Grade 6 *LPT* Mathematics			.54	.56

*The full tables of correlation statistics may be found in Appendix A.

Test Reliability Data Is Solid

Data measuring SOL test score reliability are essential to knowing whether the tests are fair and accurate measures of a student's knowledge and skills. Test reliability statistics address the degree to which the results of a test are dependable and consistently measure particular knowledge. Because Virginia's SOL test results impact student graduation and school accreditation, a high degree of reliability is critically important.

In developing the Virginia SOL tests, the developers used Kuder-Richardson Formula #20, or the KR-20, as the statistical measure of test reliability for all SOL tests except English: Writing, where person separation reliability was used. The Kuder-Richardson is a traditional procedure designed to determine the degree to which the test questions consistently measure the same body of content and skills. KR-20 values range from 0 to .99. A different writing reliability measure is used because the test contains both multiple-choice items and a student writing response. Test developers aim for a test's KR-20 value to be as high as possible, while knowing that reaching .99 is impossible. KR-20 values ranged from a low of .80 on Grade 5 History and Social Science to a high of .92 on Grade 8 Mathematics.

> *All the reliability estimates were high and were consistent with those typically obtained in their first year of a new testing program. The general rule of thumb for high-stakes decisions about individuals is a minimum of .85. Nearly all of the values presented met this criterion and those that did not were very close. Reliabilities for English and Mathematics were consistently high and all of the End-of-Course tests met or exceeded the rule of thumb criterion.*
>
> *Overall, the reliability profile for the Spring 1998 Administration of the Virginia …tests is very good for a first administration. Based on the experiences of other states, I would expect further improvements in subsequent administrations.*
>
> Dr. S. E. Phillips, Professor, Michigan State University
> Regarding the Virginia SOL testing program

> *The internal consistency evidence of reliability of total test scores is very strong. This kind of evidence is appropriate given the nature of the inferences that generalize from the test scores to the larger domain of the SOLs covered. The reliability coefficients above .80 are sufficiently high to justify use of the test scores as one source of evidence concerning the knowledge and skills of individual students on the SOLs.*
>
> Dr. James McMillan, Professor
> Virginia Commonwealth University

> *For a first-time test administration in a new testing program, a commonly accepted, professional criterion aims for reliability coefficients in the*

mid-80's. Most of the KR-20 reliability coefficients for each of the four grade levels assessed exceed this criterion and those that do not, approach it. It should be expected that these coefficients would increase with each additional test administration.

Dr. Tonya Moon
University of Virginia

Dr. William A. Mehrens, a professor at Michigan State who reviewed the reliability information, noted that the reliability data for the high school tests were "sufficiently high."

The reliability statistics from the Spring 1998 administration of the SOL tests were released on September 21, 1998. The full text of this release as well as the table of reliability coefficients may be found in Appendix B.

CONCLUSION

Virginia has provided substantial evidence that its SOL *tests are both valid and reliable. These data indicate that the* Virginia SOL *test scores accurately reflect the performance of students and schools on Virginia's new standards. The relatively low performance for the first administration is a reflection of the difficulty of the standards and the multiyear time frame needed for full implementation of the new standards in school curricula. If Virginia follows the pattern in other states, scores will increase substantially in the next few years as schools and students become better prepared for the tested content.*

Dr. S. E. Phillips, Professor, Michigan State University
Regarding the Virginia SOL testing program

Virginia's SOL tests are developed using procedures accepted within the testing industry for use in high-stakes test construction. Reliability and validity information from the first administration has, in many areas, exceeded the typical requirements for such programs. This is attributable, in large part, to the dedication of the considerable number of Virginia educators who have contributed their expertise and time to ensure that the fundamental tenets of good test construction are upheld in developing the SOL tests.

Virginia utilizes a very open and inclusive process for test development. Using Content Review Committees composed of Virginia educators expert in the content of the SOLs, in conjunction with the technical expertise contributed by the test contractor and the Department of Education, high-quality tests are developed. Virginia will continue to use these procedures to ensure that each SOL test administered meets these technical requirements.

The notion that schools that have not done well on SOL tests because the tests are faulty, invalid, or unreliable is not borne out in the data. It much more likely means that the expectations for students and schools is much higher than the national status quo as reflected by scores on the *Stanford 9*. The data indicate that expectations for Virginia's high school graduates have certainly been raised over those embodied in Virginia's grade 6 *Literacy Passport Test*.

APPENDIX A

Table A
Pass Rates on the Grade 3 *Standards of Learning (SOL)* Tests
Correlated with
National Percentile Ranks on the Grade 3 *Stanford 9* Tests

Grade 3 SOL Test (Spring 1998)/ Grade 3 *Stanford 9* Test (Spring 1997)	Spearman Rank Order Correlation Coefficients (Number of Schools)
SOL English: Reading and Writing with *Stanford 9* Reading Vocabulary	.76 (1071)
SOL English: Reading and Writing with *Stanford 9* Reading Comprehension	.77 (1071)
SOL English: Reading and Writing with *Stanford 9* Total Reading	.78 (1071)
SOL English: Reading and Writing with *Stanford 9* Language	.72 (1071)
SOL Mathematics with *Stanford 9* Mathematics: Procedures	.67 (1071)
SOL Mathematics with *Stanford 9* Mathematics: Problem Solving	.76 (1071)
SOL Mathematics with *Stanford 9* Total Mathematics	.75 (1071)

APPENDIX A
(continued)

Table B
Pass Rates on the Grade 5 *Standards of Learning (SOL)* Tests
Correlated with
National Percentile Ranks on the Grade 5 *Stanford 9* Tests

Grade 5 SOL Test (Spring 1998)/ Grade 5 *Stanford 9* Test (Spring 1997)	Spearman Rank Order Correlation Coefficients (Number of Schools)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Vocabulary	.76 (1039)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Comprehension	.77 (1039)
SOL English: Reading/Literature and Research with *Stanford 9* Total Reading	.78 (1039)
SOL English: Writing with *Stanford 9* Language	.76 (1039)
SOL Mathematics with *Stanford 9* Mathematics: Procedures	.67 (1039)
SOL Mathematics with *Stanford 9* Mathematics: Problem Solving	.76 (1039)
SOL Mathematics with *Stanford 9* Total Mathematics	.74 (1039)

APPENDIX A
(continued)

Table C
Pass Rates on the Grade 8 *Standards of Learning (SOL)* Tests
Correlated with
National Percentile Ranks on the Grade 8 *Stanford 9* Tests

Grade 8 SOL Test (Spring 1998)/ Grade 8 *Stanford 9* Test (Spring 1997)	Spearman Rank Order Correlation Coefficients (Number of Schools)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Vocabulary	.80 (368)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Comprehension	.80 (368)
SOL English: Reading/Literature and Research with *Stanford 9* Total Reading	.81 (368)
SOL English: Writing with *Stanford 9* Language	.82 (369)
SOL Mathematics with *Stanford 9* Mathematics: Procedures	.77 (368)
SOL Mathematics with *Stanford 9* Mathematics: Problem Solving	.85 (368)
SOL Mathematics with *Stanford 9* Total Mathematics	.83 (368)

APPENDIX A
(continued)

Table D
Pass Rates on Certain High School *Standards of Learning (SOL)* Tests
Correlated with
National Percentile Ranks on the Grade 11 *Stanford 9* Tests

Certain High School SOL Test (Spring 1998)/ Grade 11 *Stanford 9* Test (Spring 1997)	Spearman RankOrder Correlation Coefficients (Number of Schools)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Vocabulary	.57 (315)
SOL English: Reading/Literature and Research with *Stanford 9* Reading Comprehension	.64 (315)
SOL English: Reading/Literature and Research with *Stanford 9* Total Reading	.62 (315)
SOL English: Writing with *Stanford 9* Language	.71 (313)
SOL Algebra I with *Stanford 9* Total Mathematics	.53 (312)
SOL Geometry with *Stanford 9* Total Mathematics	.71 (308)
SOL Algebra II with *Stanford 9* Total Mathematics	.66 (307)

APPENDIX A
(continued)

Table E
Pass Rates on the Grades 5 and 8 *Standards of Learning* Tests
Correlated with
National Percentile Ranks on the Grade 6 *Literacy Passport Tests (LPT)*

Grades 5 and 8 SOL Tests (Spring 1998)/ Grade 6 *Literacy Passport Tests (LPT)* (Spring 1998)	Spearman Rank Order Correlation Coefficients (Number of Schools)
SOL Grade 5 English: Reading/Literature and Research with *LPT* Reading (Grade 6)	.64 (272)
SOL Grade 5 English: Writing with *LPT* Reading (Grade 6)	.68 (270)
SOL Grade 5 Mathematics with *LPT* Reading (Grade 6)	.54 (272)
SOL Grade 8 English: Reading/Literature and Research with *LPT* Reading (Grade 6)	.75 (288)
SOL Grade 8 English: Writing with *LPT* Reading (Grade 6)	.61 (287)
SOL Grade 8 Mathematics with *LPT* Reading (Grade 6)	.56 (290)

APPENDIX B

Press Release
September 21, 1998

NEW SOL TESTS SCORE WELL ON RELIABILITY

Virginia's new Standards of Learning tests have scored well on various technical measures of test reliability, according to an analysis conducted by outside testing experts.

The state Department of Education today released the first statistical results measuring test reliability based on the administration of the Standards of Learning (SOL) tests last spring.

"We were especially pleased with the reliability statistics from this first round of tests," said Cameron Harris, Assistant Superintendent for Assessment and Reporting. "As we developed the tests we felt we were moving in the right direction to produce solid assessments. These first numbers measuring test reliability confirm this."

While a technical topic, data measuring the SOL tests' reliability are essential to knowing whether the tests are fair and accurate measures of a student's knowledge and skills. According to the Department of Education, test reliability statistics address the degree to which the results of a test are dependable and consistently measure particular student knowledge. Since Virginia's SOL test results impact student graduation and school accreditation, a high degree of reliability is critically important.

In developing the Virginia SOL tests the developers used the Kuder-Richardson 20, or KR-20, as the statistical measure of test reliability for all SOL tests except writing where person separation reliability was used. The Kuder-Richardson is a traditional measure designed to measure the degree to which the test questions consistently measure the same body of content and skills. KR-20 values range from 0 to .99. A different writing reliability measure is used since the test contains both multiple choice items and a student writing response. Test developers aim for a test's KR-20 value to be as high as possible, knowing that reaching .99 is impossible.

"Our KR-20 values ranged from a low of .80 on Grade 5 History and Social Science to a high of .92 on Grade 8 Mathematics. In my experience, these numbers are

remarkably high for a first-time test and are very comparable to those of other tests that have been given for a much longer time, such as the Stanford tests. We are especially pleased that our highest reliability scores are on the high school tests which have the highest stakes for individual students. " Harris noted.

The Virginia SOL reliability profile was reviewed by two outside testing experts. Dr. Susan E. Phillips, Professor at Michigan State University, noted that "All the reliability estimates were high and were consistent with those typically obtained in their first year of a new testing program. The general rule of thumb for high-stakes decisions about individuals is a minimum of 0.85. Nearly all of the values presented met this criterion and those that did not were very close. Reliabilities for English and Mathematics were consistently high and all of the End-of-Course tests met or exceeded the rule of thumb criterion."

> Dr. William A. Mehrens, also a professor at Michigan State who reviewed the information, noted when referring to the high school tests, that the reliability data were "sufficiently high."

Dr. Phillips concluded, "Overall, the reliability profile for the Spring 1998 Administration of the Virginia …tests is very good for a first administration. Based on the experiences of other states, I would expect further improvements in subsequent administrations."

"It is great that the confidence we have put in our tests has been confirmed by the analysis of outside testing experts." said Kirk T. Schroder, President of the Virginia Board of Education. "As we arrive at the point to set passing scores, it is very reassuring to know that our tests have such solid reliability."

# # #

Virginia Standards of Learning
Spring 1998 Administration

Kuder-Richardson #20 Reliability Coefficients

Grade Three SOL Tests

SOL Test	Number of Test Questions	KR #20
English	45	0.90
Mathematics	50	0.91
History and Social Science	40	0.84
Science	40	0.85

Grade Five SOL Tests

SOL Test	Number of Test Questions	KR #20
English: Reading/Literature & Research	42	0.89
Mathematics	50	0.88
History and Social Science	40	0.80
Science	40	0.81
Computer/Technology	30	0.81
English: Writing*	21*	0.84**

* 20 multiple-choice items and 1 writing prompt   ** person separation reliability

Grade Eight SOL Tests

SOL Test	Number of Test Questions	KR #20
English: Reading/Literature & Research	42	0.87
Mathematics	60	0.92
History and Social Science	50	0.88
Science	50	0.88
Computer/Technology	40	0.86
English: Writing*	21*	0.82**

* 20 multiple-choice items and 1 writing prompt   ** person separation reliability

Virginia Standards of Learning
Spring 1998 Administration
Kuder-Richardson #20 Reliability Coefficients

High School SOL Tests

SOL Test	Number of Test Questions	KR #20
English: Reading/Literature & Research	42	0.89
Algebra I	50	0.88
Geometry	45	0.85
Algebra II	50	0.86
U. S. History	61	0.90
World History to 1000 + World Geography	61	0.91
World History from 1000 + World Geography	63	0.91
Biology	50	0.88
Earth Science	50	0.87
Chemistry	50	0.88
Writing	31*	0.86**

* 30 multiple-choice items and 1 writing prompt   ** person separation reliability